OCR PROJECT

ADVISOR- DR. SULA ARDIANA
DR. AMINUL ISLAM MUHAMMAD

University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING

Introduction

Overview of the project: This project focuses on automating the extraction of key information from transcripts submitted by applicants to University of New Haven.

Objective: The primary objective is to streamline the review process for new applications by automating the extraction and organization of important information such as student names, universities, courses/subjects, grades, and GPAs.

Target Audience: The project is designed to benefit the Admissions Committee, making their review process more efficient and effective.

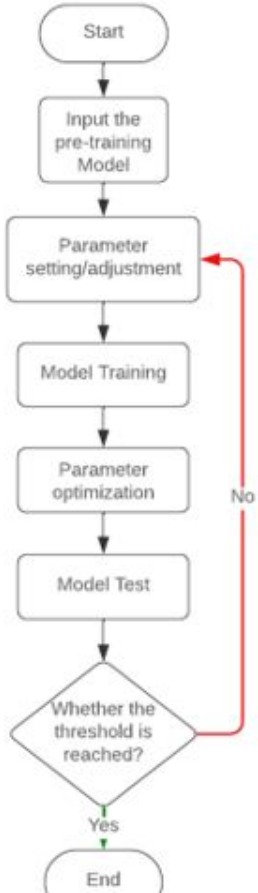
Benefits: By automating this process, we aim to increase efficiency, reduce errors, and provide a more streamlined and organized approach to reviewing applications.

OCR Project RoadMap

OCR-App: Enhancement and Training of Transcript OCR Model and GPA Prediction Model

- FA 23:
  - The project utilized OCR tools such as Tesseract and PyPDF, and experimented with several models, including Microsoft Table Transformer, Paddle OCR, and EasyOCR, among others.
- SP24:
  - **Optimize OCR Accuracy** - Enhance ability to process diverse transcript formats.
  - **Enhance Data Extraction and Feature Engineering** - Improve data extraction for GPA prediction.
  - **Extend Model Training Across Indian Universities** - Broaden training data for better generalizability.
  - **Improve Data Output Quality** - Align CSV output with data analysis needs.
  - **Advance GPA Prediction Model** - Enhance GPA predictions for data-driven admissions.

Fine-tuning PaddleOCR for Transcript Analysi



Customization Process:

- Preparing the custom transcripts dataset for fine-tuning.
- Configuring the PaddleOCR model for fine-tuning.
- Training the model on the custom dataset.

Benefits of Fine-tuning:

- Improved accuracy in extracting information from transcripts.
- Adaptation to the specific characteristics of our transcripts, such as font styles and layouts.

Challenges in Fine-tuning and Post-processing:

Dataset Preparation:

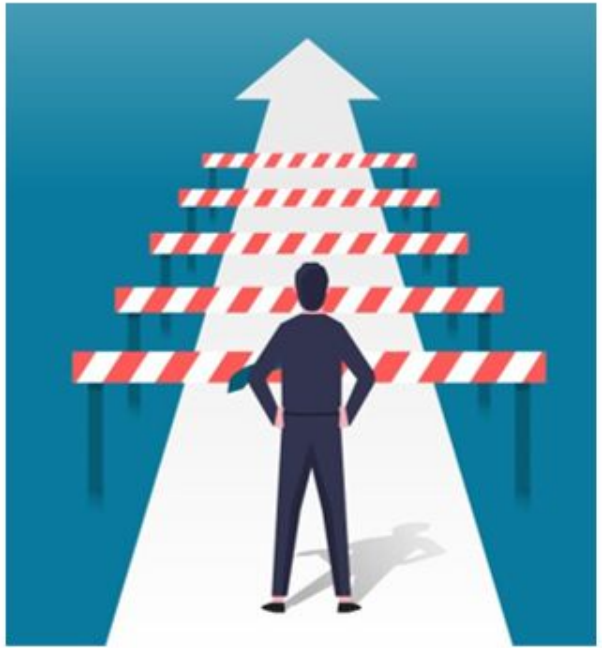
- Ensuring the dataset is representative of the transcripts to be processed.
- Handling variations in formatting, layout, and quality of transcripts.

Fine-tuning Issues:

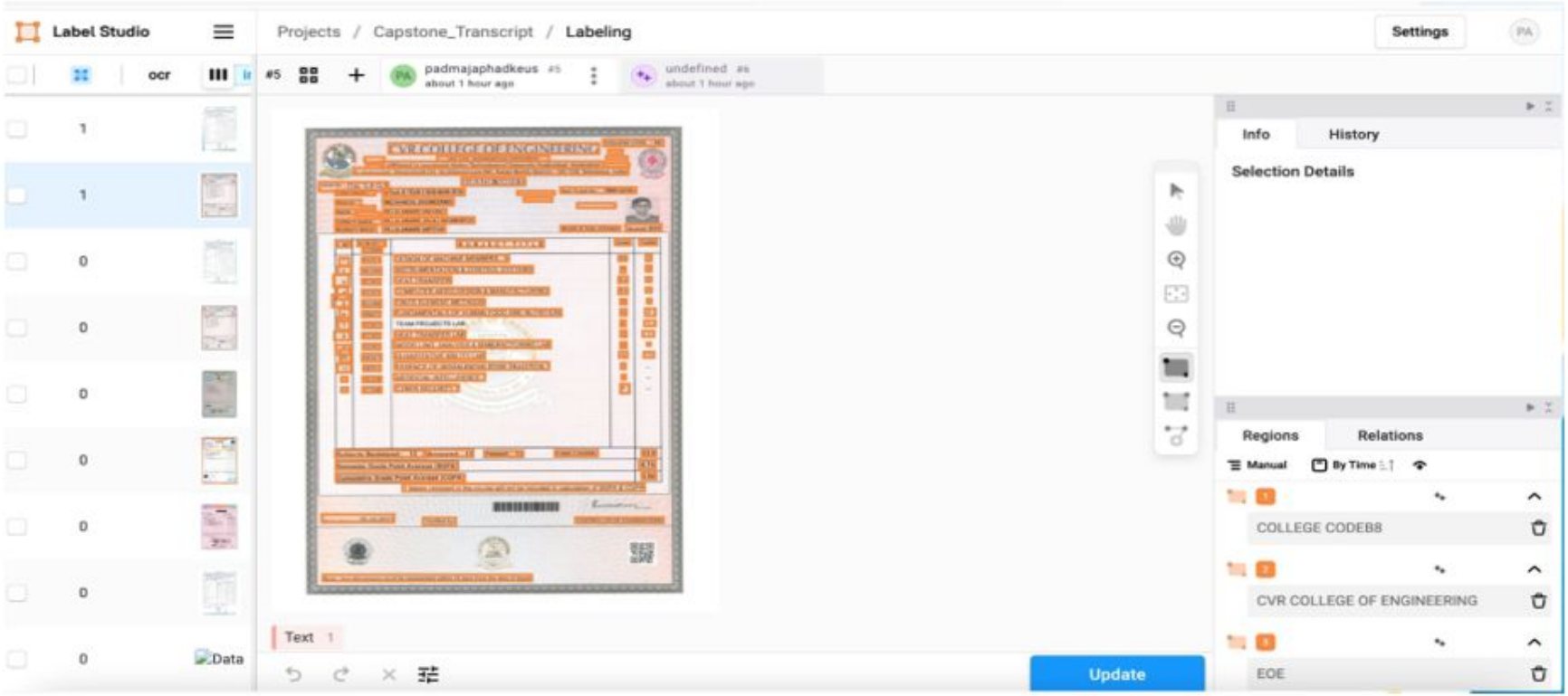
- Finding the right balance between underfitting and overfitting.
- Addressing domain-specific challenges in transcript data.

Post-processing Challenges:

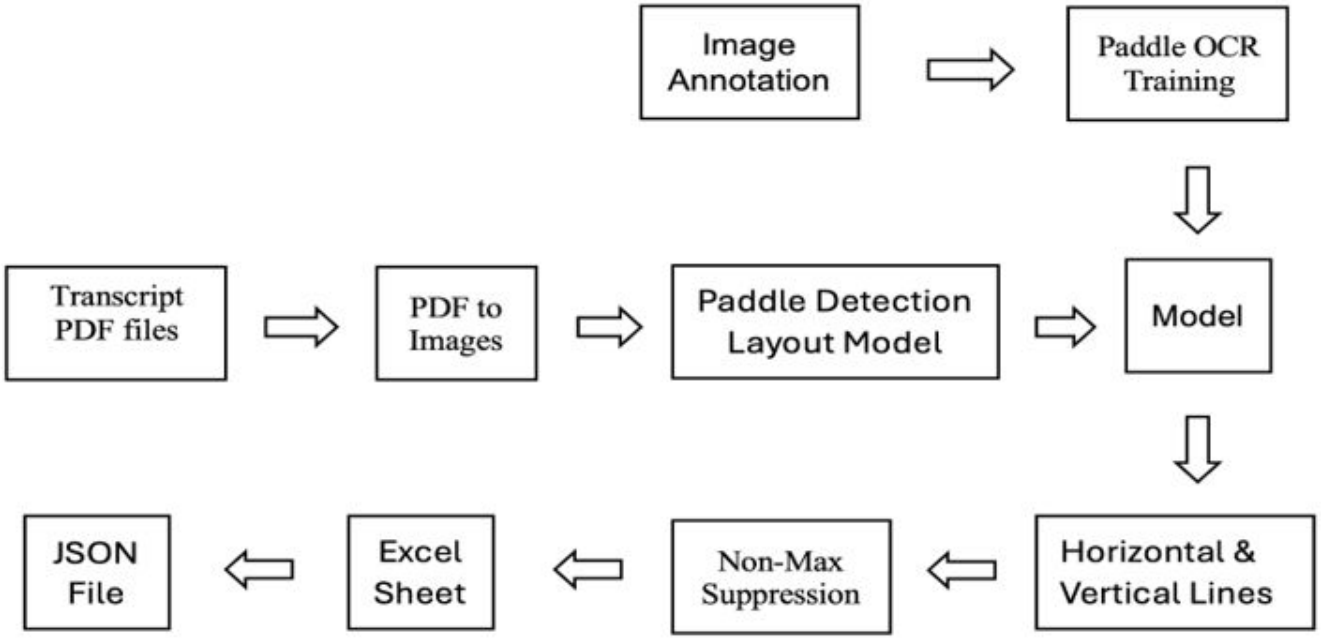
- Ensuring the extracted information is accurate and formatted correctly.
- Handling variations in how information is presented in transcripts.



Label Studio Annotation



Block Diagram



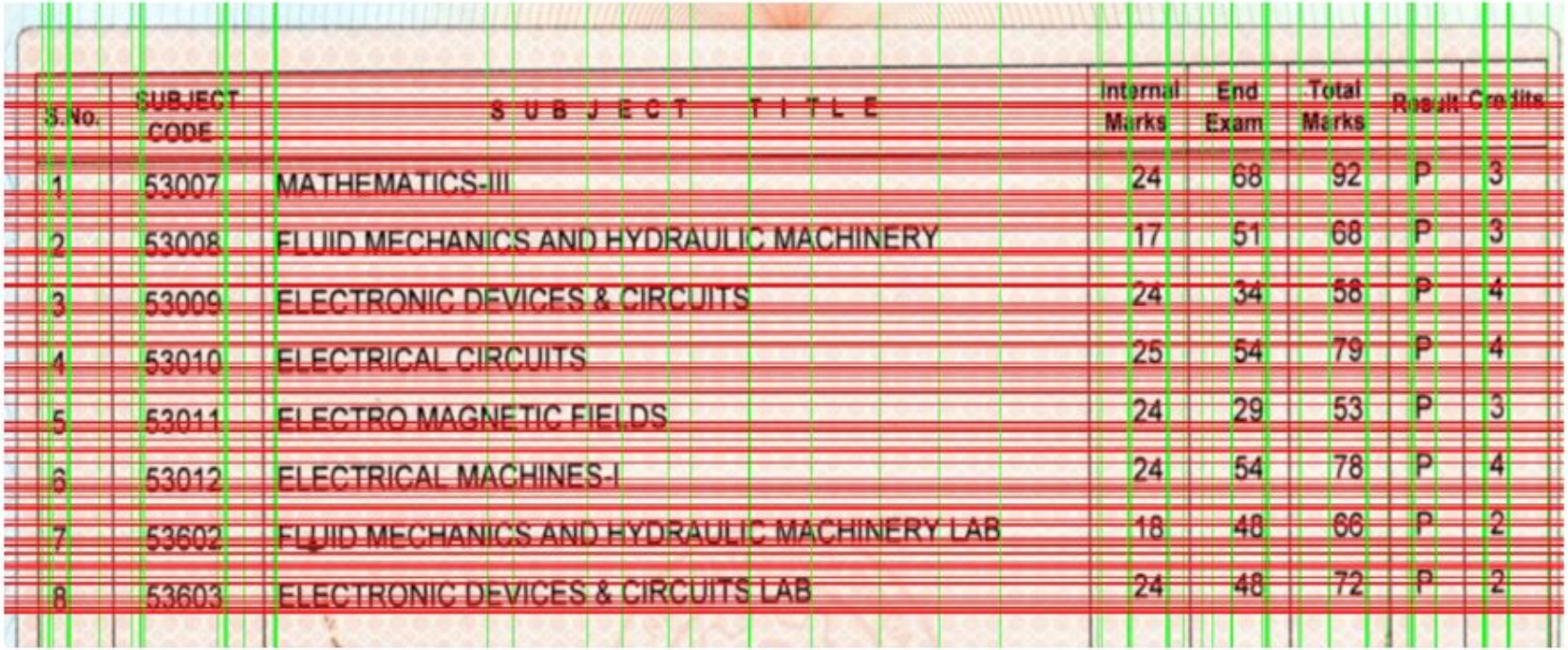
Postprocessing Techniques

- Text Detection and Recognition
- Horizontal and Vertical Lines
- Non-Max Suppression
- Conversion to Excel sheet
- Handling Duplicates
- Conversion to JSON File

Text Detection and Recognition

S.No.	SUBJECT CODE	SUBJECT TITLE	Internal Marks	End Exam	Total Marks	Result	Credits
1	53007	MATHEMATICS-III	24	68	92	P	3
2	53008	FLUID MECHANICS AND HYDRAULIC MACHINERY	17	51	68	P	3
3	53009	ELECTRONIC DEVICES & CIRCUITS	24	34	58	P	4
4	53010	ELECTRICAL CIRCUITS	25	54	79	P	4
5	53011	ELECTRO MAGNETIC FIELDS	24	29	53	P	3
6	53012	ELECTRICAL MACHINES-I	24	54	78	P	4
7	53602	FLUID MECHANICS AND HYDRAULIC MACHINERY LAB	18	48	66	P	2
8	53603	ELECTRONIC DEVICES & CIRCUITS LAB	24	48	72	P	2

Horizontal and Vertical Lines



S.No.	SUBJECT CODE	S U B J E C T T I T L E	Internal Marks	End Exam	Total Marks	Result	Credits
1	53007	MATHEMATICS-III	24	68	92	P	3
2	53008	FLUID MECHANICS AND HYDRAULIC MACHINERY	17	51	68	P	3
3	53009	ELECTRONIC DEVICES & CIRCUITS	24	34	58	P	4
4	53010	ELECTRICAL CIRCUITS	25	54	79	P	4
5	53011	ELECTRO MAGNETIC FIELDS	24	29	53	P	3
6	53012	ELECTRICAL MACHINES-I	24	54	78	P	4
7	53602	FLUID MECHANICS AND HYDRAULIC MACHINERY LAB	18	48	66	P	2
8	53603	ELECTRONIC DEVICES & CIRCUITS LAB	24	48	72	P	2

Non-Max Suppression

S.No.	SUBJECT CODE	SUBJECT TITLE	Internal Marks	End Exam	Total Marks	Result	Credits
1	53007	MATHEMATICS-III	24	68	92	P	3
2	53008	FLUID MECHANICS AND HYDRAULIC MACHINERY	17	51	68	P	3
3	53009	ELECTRONIC DEVICES & CIRCUITS	24	34	58	P	4
4	53010	ELECTRICAL CIRCUITS	25	54	79	P	4
5	53011	ELECTRO MAGNETIC FIELDS	24	29	53	P	3
6	53012	ELECTRICAL MACHINES-I	24	54	78	P	4
7	53602	FLUID MECHANICS AND HYDRAULIC MACHINERY LAB	18	48	66	P	2
8	53603	ELECTRONIC DEVICES & CIRCUITS LAB	24	48	72	P	2

Excel File

S.No.	SUBJECT CODE	SUBJECT TITLE	Internal Marks	End Exam	Total Marks	Resuit	Credits
1	53007	MATHEMATICS-III	24	68	92	P	3
2	53008	FLUID MECHANICS AND HYDRAULIC MACHINERY	17	51	68	P	3
3	53009	ELECTRONIC DEVICES & CIRCUITS	24	34	58	P	4
4	53010	ELECTRICAL CIRCUITS	25	54	79	P	4
5	53011	ELECTRO MAGNETIC FIELDS	24	29	53	P	3
	53012	ELECTRICAL MACHINES-I	24	54	78	P	4
7	53602	FLUID MECHANICS ANDHYDRAULIC MACHINERY LAB	18	48	66	P	2
8	53603	ELECTRONIC DEVICES & CIRCUITS LAB	24	48	72	P	2

JSON File

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD
KUKATPALLY, HYDERABAD - 500 085
ANDHRA PRADESH, INDIA
MEMORANDUM OF MARKS

M 3816416
11042006390
II B.Tech. I Semester (R09) Regular
ELECTRICAL & ELECTRONICS ENGINEERING
S PRAVEENA KUMARI
10291A0234
November, 2011
29-V R E C, NIZAMABAD

S.No.	SUBJECT CODE	SUBJECT TITLE	Internal Marks	End Exam	Total Marks	Result	Credits
1	53007	MATHEMATICS-III	24	68	92	P	3
2	53008	FLUID MECHANICS AND HYDRAULIC MACHINERY	17	51	68	P	3
3	53009	ELECTRONIC DEVICES & CIRCUITS	24	34	58	P	4
4	53010	ELECTRICAL CIRCUITS	25	54	79	P	4
5	53011	ELECTRO MAGNETIC FIELDS	24	29	53	P	3
6	53012	ELECTRICAL MACHINES-I	24	54	78	P	4
7	53602	FLUID MECHANICS AND HYDRAULIC MACHINERY LAB	18	48	66	P	2
8	53603	ELECTRONIC DEVICES & CIRCUITS LAB	24	48	72	P	2
SUBJECTS REGISTERED: 8		APPEARED: 8	PASSED: 8	TOTAL:	160	386	566		25

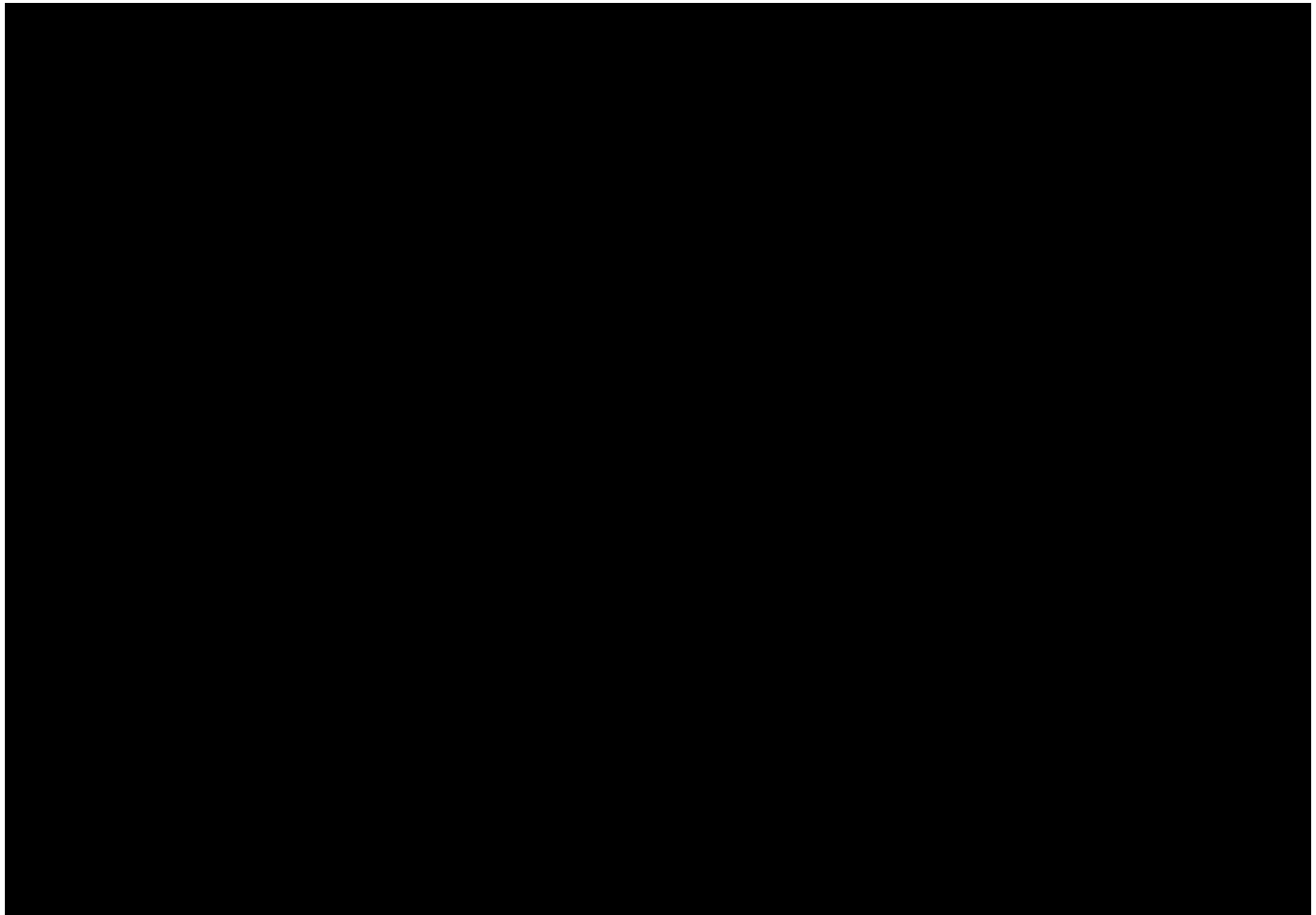
Aggregate (in Words) **** Five Six Six ****
Date of Issue : February 27, 2012 Verified by
CONTROLLER OF EXAMINATIONS

Instructions :

	Maximum Marks			Minimum Marks for pass	
	Internal	End Exam.	Total of Int. & End	End Exam.	Total of Int. & End
Theory Subjects	25	75	100	26	40
Practical Subjects	25	50	75	18	30

Note : Any discrepancy must be represented within 15 days from the date mentioned above. P: PASS F: FAIL AB: ABSENT


```
{
  "Semester 1": {
    "Name": "SPRAVEENA KUMARI",
    "University": "JAWAHARLALNEHRU TECHNOLOGICAL UNIVERSITYHYDERABAD",
    "Course": "ELECTRICAL & ELECTRONICS ENGINEERING",
    "CGPA": NaN,
    "Percentage": NaN,
    "Autonomous": NaN,
    "Course Info": [
      {
        "S.No.": 1,
        "SUBJECT CODE": 53007,
        "SUBJECT TITLE": "MATHEMATICS-III",
        "Internal Marks": 24,
        "End Exam": 68,
        ".Total Marks": 92,
        "Result Credits": "P",
        "Unnamed: 7": 3
      },
      {
        "S.No.": 2,
        "SUBJECT CODE": 53008,
        "SUBJECT TITLE": "FLUID MECHANICS AND HYDRAULIC MACHINERY",
        "Internal Marks": 17,
        "End Exam": 51,
        ".Total Marks": 68,
        "Result Credits": "P",
        "Unnamed: 7": 3
      },
      {
        "S.No.": 3,
        "SUBJECT CODE": 53009,
        "SUBJECT TITLE": "ELECTRONIC DEVICES & CIRCUITS",
        "Internal Marks": 24,
        "End Exam": 34,
        ".Total Marks": 58,
        "Result Credits": "P",
        "Unnamed: 7": 4
      },
```



THANK YOU!